ISSALE 2014 – Course 1B

Resources and Methods in Corpus-Based Lexical Semantics

— Week 2 —

Prof. Dr. Stefan Evert

FAU Erlangen-Nürnberg

www.stefan-evert.de

Suggested project topics

- Sentiment analysis
  - data: SemEval 2013 Twitter
  - generate polarity lexica
- WSD
  - data: SemCor (NLTK)
  - Lesk vs. ML vs. distributional
- Co-occurrence data
  - provided corpus or own data
  - DSM, collocations, keywords, selectional preferences
- Topic clustering
  - latent semantic indexing vs. context vectors vs. LDA
- Web GUIs with R + Shiny

Mini lectures (planned)

- WordNet & WSD
  - with Python + NLTK
- Sentiment analysis
  - polarity classification
- Corpus indexing & search
  - indexing annotated corpora
  - pattern-based search (regexp)
- Collocations & keywords
  - also: selectional preferences
- Matrix algebra in R
  - implementation of DSM, dimensionality reduction
  - topic clustering
- N-gram databases

Software & data

- Python
  - NLTK, Lesk algorithm
  - scikit-learn, Rpy2
  - polarity classifier
- IMS Corpus Workbench
  - with Perl API
  - experimental Python API
  - CQPweb GUI (online)
- UCS toolkit
  - Perl/R
  - cooccurrence extraction
- Sentiment analysis
  - SemEval 2013 + 2014
  - Web: movie reviews etc.
- WSD
  - SemCor (NLTK)
  - Web: SemEval/SensEval
- Corpora
  - BNC, Gigaword, Web
  - movie subtitles, Twitter, financial disclosures, …
- N-Grams
  - Google Web1T5 (huge!)
  - Twitter N-grams

Software installation (1)

- Update Ubuntu (just in case)
  - Update Manager → Check → Install Updates
- We can't get shared folders to work ☹
- Install additional virtual hard disk
  - copy disk image & follow instructions from server
  - need to mount after every reboot of Ubuntu VM:
    `sudo mount /dev/sdb1 ~/Desktop/corpus`
  - make writable for your user (once):
    `sudo chown -R ucsc ~/Desktop/corpus`
  - you can also mount in a more convenient location, but the following instructions assume `~/Desktop/corpus`

Software installation (2)

- Download data files from server
  - open file browser (“Nautilus”)
  - press Ctrl+K, then enter URL
    `smb://192.168.200.245/issaledemo`
    and your server password
  - you'll find a basic set of corpora and other resources in the directory “`corpas`” (sic!)
  - please copy only the data you need (it's very slow)
  - additional data may be uploaded later / on USB pen drive

Software installation (3)

- Create directory for software installation
  - `mkdir ~/Desktop/Software`
  - `cd ~/Desktop/Software`

- Python (2.x or 3.x)
  - it's easier to stick with Python 2.7 for now
  - `sudo apt-get install python-numpy python-scipy python-pip`
  - `sudo pip install -U scikit-learn PyStemmer regex rpy2`

Software installation (4)

- NLTK (version 3.0 *sigh*)
  - installation guide: http://nltk.org/
  - `sudo pip install -U nltk`
- NLTK corpora & data sets
  - `mkdir ~/Desktop/corpus/nltk_data`
  - `sudo ln -s ~/Desktop/corpus/nltk_data /usr/local/share`
  - now install corpora required for today's session:
    `python -m nltk.downloader wordnet wordnet_ic semcor`
  - specify `all` to download all data sets (1.9 GB!!)
  - you can also copy individual subdirectories from our server

Software installation (5)

- IMS Corpus Workbench (3.4 beta)
  - http://cwb.sf.net/developers.php
  - `sudo apt-get install subversion`
  - then check out source code from SVN:
    `svn co http://svn.code.sf.net/p/cwb/code/cwb/trunk cwb`
  - compile source code and install
  - `cd cwb`
  - `sudo SITE=standard install-scripts/cwb-install-ubuntu`
  - `cd ..`

Software installation (6)

- Installing indexed corpora
  - prepare central "registry" for CWB-indexed corpora
  - `mkdir -p ~/Desktop/corpus/CWB/registry`
  - `sudo mkdir /usr/local/share/cwb`
  - `sudo ln -s ~/Desktop/corpus/CWB/registry /usr/local/share/cwb`
  - install index files anywhere you like
  - copy registry file to `~/Desktop/corpus/CWB/registry`
  - edit file paths in `HOME` and `INFO` fields
- We will install one or two sample corpora together

Software installation (7)

- CWB/Perl API (for 3.4 beta)
  - `cd ~/Desktop/Software`
  - check out API source code from SVN:
    `svn co http://svn.code.sf.net/p/cwb/code/perl/trunk cwb-perl`
  - in subdirectories `CWB` and `CWB-CL`, type these commands:
  - `perl Makefile.PL`
  - `make test`
  - `sudo make install`
  - `make clean`

Software installation (8)

- UCS toolkit (cutting edge)
  - http://www.collocations.de/software.html
  - `cd ~/Desktop/Software`
  - check out cutting-edge version from SVN:
    ```
    svn co svn://svn.code.sf.net/p/multiword/
    code/software/UCS/trunk UCS
    ```
  - install required packages:
    ```
    sudo apt-get install libexpect-perl
         libterm-readkey-perl a2ps
    ```

Software installation (9)

- Configure UCS toolkit
  - `cd UCS/System`
  - `perl Install.perl`
  - now link command-line tools into search path
  - `rm bin/*~`
  - `sudo ln -s `pwd`/bin/ucs* /usr/local/bin`
- A quick test (optional)
  - compute association scores for given contingency table:
    `ucs-list-am -f 12,100,2000,99999`
  - most frequent adjective-noun pairs in Dickens corpus:

```
ucs-sort dickens.ds.gz by f- l2 l1
    | ucs-print -i
```

We're done!